STABILITY OF FINITE DIFFERENCE SCHEMES FOR HYPERBOLIC INITIAL BOUNDARY VALUE PROBLEMS: NUMERICAL BOUNDARY LAYERS.

BENJAMIN BOUTIN & JEAN-FRANÇOIS COULOMBEL

Abstract. In this article, we give a unified theory for constructing boundary layer expansions for discretized transport equations with homogeneous Dirichlet boundary conditions. We exhibit a natural assumption on the discretization under which the numerical solution can be written approximately as a two-scale boundary layer expansion. In particular, this expansion yields discrete semigroup estimates that are compatible with the continuous semigroup estimates in the limit where the space and time steps tend to zero. The novelty of our approach is to cover numerical schemes with arbitrarily many time levels, while semigroup estimates were restricted, up to now, to numerical schemes with two time levels only.

AMS classification: 65M12, 65M06, 65M20.

Keywords: transport equations, numerical schemes, Dirichlet boundary condition, boundary layers, stability.

Contents

1. Introduction and main result

1.1. Introduction.

The analysis of numerical boundary conditions for hyperbolic equations is a delicate subject for which several definitions of stability can be adopted. Any such definition relies on the choice of a given topology that is a discrete analogue of the norm of some functional space in which the underlying continuous problem is known to be well-posed. The stability theory for numerical boundary conditions developed in [GKS72], though rather natural in view of the results of [Kre70] for partial differential equations, may have suffered from its "technicality". As Trefethen and Embree [TE05, chapter 34]

Date: February 10, 2015.

Research of B. B. was partially supported by ANR project ACHYLLES, ANR-14-CE25-0001-03.

Research of J.-F. C. was partially supported by ANR project BoND, ANR-13-BS01-0009-01.

say: "[...] *the term GKS-stable is quite complicated. This is a special definition of stability,* [...], *that involves exponential decay factors with respect to time and other algebraic terms that remove it significantly from the more familiar notion of bounded norms of powers*". More precisely, the definition of stability in [GKS72] corresponds to norms of $\ell^2_{t,x}$ type for the numerical solution (t denotes time and x denotes the space variable), while in many problems of evolutionary type one is more used to the $\ell^\infty_t(\ell^2_x)$ topology. In terms of operator theory, the definition of stability in [GKS72] corresponds to *resolvent estimates*, while the more familiar notion of bounded norms of powers corresponds to *semigroup estimates*. Hence a natural -though delicate- question in the theory of hyperbolic boundary value problems is to pass from GKS type (that is, resolvent) estimates to semigroup estimates. In the context of partial differential equations, this problem has received a somehow final answer in [Mét14], see references therein for historical comments on this problem. In the context of numerical schemes, the derivation of semigroup estimates is not as well understood as for partial differential equations. Semigroup estimates have been derived in [Wu95] for discrete scalar equations, and in [CG11] for systems of equations. However, the analysis in [Wu95] and [CG11] only deals with schemes with two time levels, and does not extend as such to schemes with three or more time levels (e.g., the leap-frog scheme).

In this article, we focus on Dirichlet boundary conditions and derive semigroup estimates for a class of numerical schemes with arbitrarily many time levels. The reasons why we choose Dirichlet boundary conditions are twofold. First, these are the only boundary conditions for which, independently of the (stable) numerical scheme that is used for discretizing a scalar transport equation, stability in the sense of GKS is known to hold. The latter result dates back to [GT81] and is recalled later on. Second, homogeneous Dirichlet boundary conditions typically give rise to numerical boundary layers and therefore to an accurate description of the numerical solution by means of a two-scale expansion. We combine these two favorable aspects of the Dirichlet boundary conditions in our derivation of a semigroup estimate.

The study of numerical boundary layers has received much attention in the past decades, including for nonlinear systems of conservation laws, see for instance [DL88, GS97, CHG01]. As far as we know, all previous studies have considered numerical schemes with a three point stencil and two time levels. In this article, we focus on linear transport equations and exhibit a class of numerical schemes for which the homogeneous Dirichlet boundary conditions give rise to numerical boundary layers. The stencil can be arbitrarily wide. As follows from our criterion, the occurrence of boundary layers is not linked with the order of accuracy of the numerical scheme, which is a *low frequency* property, but rather with its *high frequency* behavior. For instance, the Lax-Wendroff discretization displays numerical boundary layers when combined with Dirichlet boundary conditions (and such layers have the same width as for the Lax-Friedrichs scheme) but the leap-frog scheme does not[1], though both Lax-Wendroff and leap-frog schemes are formally of order 2.

1.2. **Notations.**

We consider a one-dimensional scalar transport equation:

$$\partial_t u + a\,\partial_x u = 0, \quad t > 0\,, x > 0\,, \tag{1.1}$$

where the velocity is $a \neq 0$. The transport equation (1.1) is supplemented with an initial condition u_0 that belongs to a functional space that is made precise later on. In the case $a > 0$, that is, if we consider an *incoming* transport equation, we also supplement (1.1) with homogeneous Dirichlet boundary condition:

$$u(0,t) = 0, \quad t > 0\,. \tag{1.2}$$

The finite difference scheme under consideration is assumed to be obtained by the so-called *method of lines*, see, e.g., [GKO95]. In other words, we start with (1.1) and first use a space discretization. The latter is supposed to be linear with r points on the left and p points on the right. In other words, we consider some coefficients $a_{-r}, \dots, a_p$, where p, r are fixed nonnegative integers, together with a space step $\Delta x > 0$, and approximate (1.1) by the system of ordinary differential equations:

$$\dot{u}_j + \frac{1}{\Delta x} \sum_{\ell=-r}^{p} a_\ell\, u_{j+\ell} = 0\,, \tag{1.3}$$

[1] The leap-frog scheme rather generates incoming highly oscillating wave packets, as explained at the end of this article.

where $u_j(t)$ represents an approximation of the solution u to (1.1) in the neighborhood of the point $x_j := j\,\Delta x$. The integers r,p are fixed by assuming $a_{-r} \neq 0$ and $a_p \neq 0$. The latter system of ordinary differential equations is then approximated by means of a (possibly multistep) explicit numerical integration method. We refer to [HNW93, HW96] for an extensive study of numerical methods for ordinary differential equations. Applying a linear explicit multistep method to (1.3) yields the numerical approximation

$$\sum_{\sigma=0}^{k} \alpha_\sigma \, u_j^{n+\sigma} + \lambda \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} a_\ell \, u_{j+\ell}^{n+\sigma} = 0\,. \tag{1.4}$$

with $k \geq 1$ and fixed constants $\alpha_0, \dots, \alpha_k, \beta_0, \dots, \beta_{k-1}$. The multistep integration method is normalized by assuming $|\alpha_0| + |\beta_0| > 0$ and $\alpha_k = 1$. In (1.4), we have made use of the notation $\lambda := \Delta t / \Delta x$ for the so-called Courant-Friedrichs-Lewy parameter. In what follows, the parameter λ is kept fixed[2], and we consider the space and time grid $x_j := j\,\Delta x$, $t^n := n\,\Delta t$ for $j,n \in \mathbb{N}$. For notational convenience, we introduce the (dimensionless) constant $\tau > 0$ that satisfies

$$\Delta x = \tau \, |a| \, \Delta t\,. \tag{1.5}$$

We keep $\Delta t \in (0,1]$ as the only small parameter and $\Delta x \in (0, 1/\lambda]$ varies accordingly.

Since we are approximating the transport equation (1.1) on the half-line $\mathbb{R}^+$, the space grid is indexed by $\mathbb{N}$. This means that the numerical approximation (1.4) takes place for $j \geq r$. We then supplement (1.4) with homogeneous Dirichlet boundary conditions on the "numerical" boundary:

$$u_j^n = 0, \quad 0 \leq j \leq r-1, \quad n \geq k\,, \tag{1.6}$$

independently of the sign of a. The scheme (1.4), (1.6) is ignited by k initial data, which correspond to the approximation of the solution to (1.1) at times $t^0, \dots, t^{k-1}$. For simplicity, we assume that the initial data for (1.4), (1.6) are given by the standard piecewise constant approximation of the exact solution to (1.1). In other words, we set:

$$u_j^n := \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} u_0(x - a\,t^n)\,\mathrm{d}x\,, \quad j \geq 0\,, \quad n = 0, \dots, k-1\,, \tag{1.7}$$

where the initial condition u_0 for (1.1) has been extended by zero to $\mathbb{R}^-$ in the case $a > 0$.

The following two assumptions are the minimal consistency requirements for the numerical scheme (1.4).

Assumption 1.1 (Consistency of the space discretization)**.** *The coefficients $a_{-r}, \dots, a_p$ in* (1.4) *satisfy*

$$\sum_{\ell=-r}^{p} a_\ell = 0\,, \tag{1.8}$$

$$\sum_{\ell=-r}^{p} \ell \, a_\ell = a\,. \tag{1.9}$$

Assumption 1.2 (Consistency of the linear multistep integration method)**.** *The coefficients $\alpha_0, \dots, \alpha_k$, $\beta_0, \dots, \beta_{k-1}$ of the time integration method in* (1.4) *satisfy*

$$\sum_{\sigma=0}^{k} \alpha_\sigma = 0\,, \quad \sum_{\sigma=0}^{k} \sigma \, \alpha_\sigma = \sum_{\sigma=0}^{k-1} \beta_\sigma\,.$$

In the case $k = 1$, that is for numerical schemes with two time levels, the normalization gives $\alpha_0 = \alpha_1 = \beta_0 = 1$, and (1.4) reduces to the standard form

$$u_j^{n+1} - u_j^n + \lambda \sum_{\ell=-r}^{p} a_\ell \, u_{j+\ell}^n = 0\,.$$

[2] This assumption could be weakened by assuming that the ratio $|a|\,\Delta t/\Delta x$ is bounded from below and from above, but we shall restrict to the more common case where the ratio is fixed for simplicity.

If $p = r = 1$, we obtain the class of three point schemes that encompasses both the Lax-Friedrichs and Lax-Wendroff scheme.

As a direct consequence of the first consistency condition (1.8), it appears that the scheme (1.4) admits a conservative form in the following sense. There exists a linear numerical flux function F with real coefficients:

$$F(v_j,\dots,v_{j+p+r-1}) := \sum_{\ell=-r}^{p-1} f_\ell \, v_{j+\ell+r}\,,$$

such that

$$\sum_{\ell=-r}^{p} a_\ell \, u_{j+\ell} = F(u_{j-r+1},\dots,u_{j+p}) - F(u_{j-r},\dots,u_{j+p-1})\,. \tag{1.10}$$

In particular, (1.4) also takes the conservative form

$$\sum_{\sigma=0}^{k} \alpha_\sigma \, u_j^{n+\sigma} + \lambda \sum_{\sigma=0}^{k-1} \beta_\sigma \left(F(u_{j-r+1}^{n+\sigma},\dots,u_{j+p}^{n+\sigma}) - F(u_{j-r}^{n+\sigma},\dots,u_{j+p-1}^{n+\sigma}) \right) = 0\,. \tag{1.11}$$

From the second consistency condition (1.9), it follows that $F(u,\dots,u) = a\,u$ for any $u \in \mathbb{R}$. This is the usual consistency property of F with the exact flux $(u \mapsto a\,u)$ of the transport equation (1.1) written as a conservation law.

Our final assumption is the standard ℓ^2-stability assumption for (1.4) when the scheme is considered on the whole real line $j \in \mathbb{Z}$:

Assumption 1.3 (Stability for the Cauchy problem)**.** *There exists a constant $C > 0$ such that, for all $\Delta t \in (0,1]$, the solution to*

$$\sum_{\sigma=0}^{k} \alpha_\sigma \, u_j^{n+\sigma} + \lambda \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} a_\ell \, u_{j+\ell}^{n+\sigma} = 0\,, \quad j \in \mathbb{Z}\,, \quad n \in \mathbb{N}\,,$$

satisties

$$\sup_{n\in\mathbb{N}} \sum_{j\in\mathbb{Z}} \Delta x \, |u_j^n|^2 \le C \sum_{\sigma=0}^{k-1} \sum_{j\in\mathbb{Z}} \Delta x \, |u_j^\sigma|^2\,.$$

As is well-known, assumption 1.3 can be rephrased thanks to Fourier analysis. More precisely, if we introduce the function $\mathcal{A}$ defined by:

$$\forall z \in \mathbb{C}\setminus\{0\}, \quad \mathcal{A}(z) = \sum_{\ell=-r}^{p} a_\ell \, z^\ell\,, \tag{1.12}$$

then applying the Fourier transform to (1.4) yields for all $\xi \in \mathbb{R}$:

$$\sum_{\sigma=0}^{k} \alpha_\sigma \, \widehat{u^{n+\sigma}}(\xi) + \lambda \sum_{\sigma=0}^{k-1} \beta_\sigma \, \mathcal{A}(\mathrm{e}^{i\,\Delta x\,\xi}) \, \widehat{u^{n+\sigma}}(\xi) = 0\,,$$

where u^n is the piecewise constant function that takes the value u_j^n on the cell $[j\,\Delta x, (j+1)\,\Delta x)$. The stability assumption 1.3 is equivalent to requiring that there exists a constant $C > 0$ such that for all $\eta \in \mathbb{R}$, and for all given $x_0,\dots,x_{k-1} \in \mathbb{C}$, the solution $(x_\sigma)_{\sigma\in\mathbb{N}}$ to the recurrence relation

$$\forall\, n \in \mathbb{N}\,, \quad \sum_{\sigma=0}^{k} \alpha_\sigma \, x_{n+\sigma} + \lambda \sum_{\sigma=0}^{k-1} \beta_\sigma \, \mathcal{A}(\mathrm{e}^{i\,\eta}) \, x_{n+\sigma} = 0\,,$$

satisfies

$$\sup_{n\in\mathbb{N}} |x_n|^2 \le C \left(|x_0|^2 + \cdots + |x_{k-1}|^2 \right).$$

In particular, the closed curve $\{-\lambda \mathcal{A}(\mathrm{e}^{i\,\eta})\,,\,\eta \in \mathbb{R}\}$ should be contained in the so-called stability region of the numerical integration method, see [HW96, Definition V.1.1]. Observe now that the consistency assumption 1.1 introduced above can be rewritten under the form:

$$\mathcal{A}(1) = 0 \quad \text{and} \quad \mathcal{A}'(1) = a \neq 0\,. \tag{1.13}$$

Since $\mathcal{A}$ vanishes at 1, 0 should belong to the stability region of the numerical integration method, which implies (see [HNW93, Chapter III.3]):

$$\sum_{\sigma=0}^{k} \sigma\,\alpha_\sigma \neq 0\,. \tag{1.14}$$

Remark 1.4. *In the case $k = 1$, the stability assumption 1.3 is equivalent to:*

$$\forall\, z \in \mathbb{S}^1\,, \quad |1 - \lambda\,\mathcal{A}(z)| \leq 1\,. \tag{1.15}$$

In particular, assumption 1.3 constraints the CFL number λ to be "small enough", and $\mathcal{A}(z)$ can not be a negative number.

1.3. Main result.

The main result of this paper is the following theorem.

Theorem 1.5 (Semigroup estimate)**.** *Consider a linear scheme of the form (1.4) satisfying the consistency assumptions 1.1 and 1.2, the stability assumption 1.3 and the "dissipative" assumption 2.1 introduced later on. Consider an initial condition $u_0 \in H^2(\mathbb{R}_+)$ for (1.1) such that*

$$\begin{cases} u_0(0) = 0\,, & \text{if } a < 0\,,\\ u_0(0) = u_0'(0) = 0\,, & \text{if } a > 0\,. \end{cases}$$

Let $T > 0$ and, for $\Delta t \in (0, 1]$, let us define N_T as the largest integer such that $\Delta t\,N_T \leq T$. Let also $\mu \in [0, 1/3]$. Then there exists a constant $C > 0$, that is independent of $T, \Delta t, \mu, u_0$ such that the solution $(u_j^n)_{j\geq 0, n\geq 0}$ to (1.4)-(1.6)-(1.7) satisfies

$$\sup_{n\leq N_T} \sum_{j\geq 0} \Delta x\,|u_j^n|^2 \leq C \left(\|u_0\|^2_{L^2(\mathbb{R}^+)} + \Delta t^{1-3\,\mu}\, \mathrm{e}^{2\,T\,\Delta t^{\mu}}\, \|u_0\|^2_{H^2(\mathbb{R}^+)} \right). \tag{1.16}$$

Let us observe that (1.16) is compatible with the "continuous" estimate

$$\sup_{t\geq 0} \|u(t)\|^2_{L^2(\mathbb{R}^+)} \leq C\,\|u_0\|^2_{L^2(\mathbb{R}^+)}\,,$$

as Δt tends to zero. The role of assumption 2.1 is to derive a boundary layer expansion for $(u_j^n)_{j\geq 0, n\geq 0}$, that is to decompose (u_j^n) as in [DL88, GS97, CHG01] under the form

$$u_j^n \sim u^{\text{int}}(x_j, t^n) + u^{\text{bl}}(j, t^n)\,,$$

where the *boundary layer profile* u^{bl} depends on the "fast" variable $j = x_j/\Delta x$ and has exponential decay at infinity, while the *interior profile* u^{int} depends on the "slow" variable x_j. As follows from the analysis below, the derivation of such two-scale expansions is not linked to any viscous behavior of (1.4) (as the scaling $x_j/\Delta x$ might suggest at first glance).

The parameter μ can diminish the T-dependence of the constants in (1.16). In particular, given any $\varepsilon > 0$ and $T > 0$, there holds

$$\sup_{n\leq N_T} \sum_{j\geq 0} \Delta x\,|u_j^n|^2 \leq C \left(\|u_0\|^2_{L^2(\mathbb{R}^+)} + 2\,\Delta t^{1-\varepsilon}\, \|u_0\|^2_{H^2(\mathbb{R}^+)} \right),$$

for Δt sufficiently small (depending on T).

Section 2 is devoted to the construction of boundary layer expansions for solutions to (1.4), (1.6). Theorem 1.5 is proved in Section 3 by means of a careful error analysis. We discuss some examples in Section 4 together with the relevance of assumption 2.1.

2. Numerical boundary layers

2.1. **Formal derivation of the boundary layer expansion.** Our first goal is to understand when the numerical solution $(u_j^n)_{j\geq 0, n\geq 0}$ of the scheme (1.4), (1.6) can be approximated by an asymptotic boundary layer expansion:

$$u_{j,n}^{\rm app} := u^{\rm int}(x_j,t^n) + u^{\rm bl}(j,t^n)\, ,\quad j\geq 0\, ,\, n\geq 0\, .$$

In the latter decomposition, we expect $u^{\rm bl}$ to have fast decay at infinity. The functions $u^{\rm int}$ and $u^{\rm bl}$ are to be defined in such a way that $(u_{j,n}^{\rm app})$ represents an accurate approximation of (u_j^n) as Δt tends to 0. Roughly speaking, the term $u^{\rm int}$ takes care of the interior behavior of the solution far from the boundary, and $u^{\rm bl}$ involves the boundary layer correction that is localized in a neighborhood of $x=0$ and matches the boundary conditions (1.6).

We shall force the approximate solution to satisfy the initial conditions:

$$u_{j,n}^{\rm app} = u_j^n\, ,\quad j\geq 0\, ,\, n=0,\dots,k-1\, . \tag{2.1}$$

In this way, the error $(u_{j,n}^{\rm app}-u_j^n)$ will satisfy a recurrence relation of the form (1.4), (1.6) with "small" source terms but will have zero initial data. We also expect the approximate solution to satisfy (1.6), or rather

$$u_{j,n}^{\rm app} \simeq 0\, ,\quad 0\leq j\leq r-1\, ,\quad n\geq k\, , \tag{2.2}$$

where, by $\simeq 0$, we mean for instance that $u_{j,n}^{\rm app}$ should be $O(\Delta t)$ on the boundary.

For technical reasons that will be made precise in Section 3, we shall define the boundary layer term through a two term expansion of the form:

$$u^{\rm bl}(j,t^n) := u^{\rm bl,0}(j,t^n) + \Delta x\, u^{\rm bl,1}(j,t^n),$$

involving a zero order term $u^{\rm bl,0}$ plus a first order corrector $u^{\rm bl,1}$ that will be used to remove part of the consistency error.

We follow the discussions in [DL88, GS97, CHG01] and briefly present hereafter a schematic derivation of the equations that will govern the three sequences $u^{\rm int}$, $u^{\rm bl,0}$ and $u^{\rm bl,1}$. To that aim, let us introduce the following consistency error:

$$\varepsilon_{j,n+k} := \frac{1}{\Delta t}\left(\sum_{\sigma=0}^{k}\alpha_\sigma\, u_{j,n+\sigma}^{\rm app} + \lambda\sum_{\sigma=0}^{k-1}\beta_\sigma\sum_{\ell=-r}^{p}a_\ell\, u_{j+\ell,n+\sigma}^{\rm app}\right)\, ,$$

with $j\geq r$, and $n\geq 0$.

- At a fixed positive distance from the boundary, the limit $\Delta t\to 0$ corresponds to $j\to+\infty$ and the boundary layer term $u^{\rm bl}$ becomes negligible with respect to $u^{\rm int}$. The above consistency error reads (up to smaller terms)

$$\varepsilon_{j,n+k} \simeq \frac{1}{\Delta t}\left(\sum_{\sigma=0}^{k}\alpha_\sigma\, u_{j,n+\sigma}^{\rm int} + \lambda\sum_{\sigma=0}^{k-1}\beta_\sigma\sum_{\ell=-r}^{p}a_\ell\, u_{j+\ell,n+\sigma}^{\rm int}\right)\, .$$

  This quantity will be of order $O(\Delta t)$ provided that $u^{\rm int}$ is a smooth solution to the continuous equation (1.1) (recall the consistency assumptions of the numerical scheme (1.4)).

- Close to the boundary, that is for a fixed index $j\geq r$, the limit $\Delta t\to 0$ makes x_j tend to zero. If the interior solution $u^{\rm int}$ is smooth enough, we get (recall that j is fixed) $u^{\rm int}(x_j,t^n) = u^{\rm int}(0,t^n)+O(\Delta t)$ and $u^{\rm int}(0,t^{n+\sigma}) = u^{\rm int}(0,t^n)+O(\Delta t)$. Then the consistency error reads[3] (up to $O(1)$ terms):

$$\varepsilon_{j,n+k} \simeq \frac{1}{\Delta t}\left(\sum_{\sigma=0}^{k}\alpha_\sigma\, u^{\rm bl,0}(j,t^{n+\sigma}) + \lambda\sum_{\sigma=0}^{k-1}\beta_\sigma\sum_{\ell=-r}^{p}a_\ell\, u^{\rm bl,0}(j+\ell,t^{n+\sigma})\right)\, .$$

[3] Here we use the consistency conditions for the coefficients in (1.4).

Due to the consistency of the numerical integration method, and assuming that $u^{\text{bl},0}$ depends smoothly enough on the time variable, we get

$$\varepsilon_{j,n+k} \simeq \frac{1}{\Delta x} \left(\sum_{\sigma=0}^{k-1} \beta_\sigma \right) \sum_{\ell=-r}^{p} a_\ell \, u^{\text{bl},0}(j+\ell, t^{n+k}) \,,$$

The first boundary layer profile $u^{\text{bl},0}$ needs therefore to satisfy the recurrence relation[4]:

$$\sum_{\ell=-r}^{p} a_\ell \, u^{\text{bl},0}(j+\ell, t^n) = 0 \,, \quad j \geq r \,, \, n \geq k \,. \tag{2.3}$$

In terms of flux quantities, the relation (2.3) corresponds to requiring

$$\sum_{\ell=-r}^{p-1} f_\ell \, u^{\text{bl},0}(j+\ell+r, t^n) \equiv \text{C}^{\text{st}} \,,$$

with an integration constant that only depends on n, but not on j. The constant is easily seen to be zero due to the required behavior of the boundary layer profiles at infinity. In addition, the boundary condition (2.2) imposes (up to an $O(\Delta t)$ term) the trace of $u^{\text{bl},0}$ on the numerical boundary:

$$u^{\text{bl},0}(j, t^n) = -u^{\text{int}}(0, t^n) \,, \quad 0 \leq j \leq r-1 \,, \, n \geq 0 \,. \tag{2.4}$$

- We still keep the index j fixed and expand the consistency error at the following order with respect to Δt. Assuming that u^{int} is smooth enough so that its associated consistency error is $O(\Delta t)$ up to the boundary, the overall consistency error reads (up to $O(\Delta t)$ terms):

$$\varepsilon_{j,n+k} \simeq \frac{1}{\Delta t} \sum_{\sigma=0}^{k} \alpha_\sigma \, u^{\text{bl},0}(j, t^{n+\sigma}) + \left(\sum_{\sigma=0}^{k-1} \beta_\sigma \right) \sum_{\ell=-r}^{p} a_\ell \, u^{\text{bl},1}(j+\ell, t^n) \,.$$

We then require the first boundary layer corrector $u^{\text{bl},1}$ to satisfy:

$$\sum_{\ell=-r}^{p} a_\ell \, u^{\text{bl},1}(j+\ell, t^n) + \frac{1}{\Delta t} \left(\sum_{\sigma=0}^{k-1} \beta_\sigma \right)^{-1} \sum_{\sigma=0}^{k} \alpha_\sigma \, u^{\text{bl},0}(j, t^{n+\sigma}) = 0 \,, \quad j \geq r \,. \tag{2.5}$$

Since our analysis considers numerical schemes of order 1 or higher, the precise value of $u^{\text{bl},1}$ on the numerical boundary does little matter since any other choice than the one below will introduce a new $O(\Delta t)$ error that will just have the same order as the interior consistency error. For simplicity, we therefore require $u^{\text{bl},1}$ to satisfy;

$$u^{\text{bl},1}(j, t^n) = 0 \,, \quad 0 \leq j \leq r-1 \,, \, n \geq 0 \,. \tag{2.6}$$

The above formal derivation of the profile equations (2.3) and (2.5) motivates the analysis of the recurrence relation (2.3). More precisely, we are going to determine the solutions to (2.3) that tend to zero at infinity. The precise definition of the approximate solution u^{app} is given in subsection 2.5.

2.2. **A preliminary result.** Let us recall that the function $\mathcal{A}$, which is linked to the amplification matrix for the scheme (1.4), is defined in (1.12). The consistency assumption 1.1 implies that 1 is a simple root of $\mathcal{A}$. The following assumption will turn out to be crucial in the forthcoming analysis.

Assumption 2.1. *The value $z = 1$ is the unique root of $\mathcal{A}$ on $\mathbb{S}^1$:*

$$\forall \, \theta \in [-\pi, \pi] \setminus \{0\} \,, \quad \mathcal{A}(\mathrm{e}^{i\,\theta}) \neq 0 \,.$$

Remark 2.2. *In the case $k = 1$, assumption 2.1 is obviously satisfied for every dissipative scheme (for which we recall that there exist $c > 0$ and $k \in \mathbb{N}^*$ such that for all $|\theta| \leq \pi$, $|1 - \lambda \mathcal{A}(\mathrm{e}^{i\,\theta})| \leq 1 - c\,\theta^{2\,k}$). However, we underline at this level that some non-dissipative schemes satisfy assumption 2.1 too, e.g. the Lax-Friedrichs scheme (that is considered in* [CHG01]*) for which $\mathcal{A}(\mathrm{e}^{i\,\theta}) = \cos\theta - 1 - i\,\lambda\,a\,\sin\theta$).*

[4] Recall that by our consistency and stability assumptions, the sum of the β_σ is nonzero.

The main result of this subsection is the following Lemma.

Lemma 2.3. *Under Assumptions 1.1, 1.2, 1.3 and 2.1, the equation $\mathcal{A}(z) = 1$ admits exactly R roots (with multiplicity) in $\mathbb{D} \setminus \{0\} = \{z \in \mathbb{C}\,,\, 0 < |z| < 1\}$ where*

$$R = \begin{cases} r, & \text{if } a < 0\,, \\ r-1, & \text{if } a > 0\,. \end{cases}$$

Let us observe that in the case $a > 0$, r can not be zero and therefore one gets a nonnegative integer for R. Indeed, the value $r = 0$ is prohibited by the fact that the numerical dependence domain would not include the "continuous" dependence domain, see [CFL28].

The proof of Lemma 2.3 makes use of the following simple observation which we have not found in [HW96] and therefore state here. We keep the notations of [HNW93, Chapter III.2].

Lemma 2.4. *Consider an ordinary differential equation of the form $\dot{y} = f(y)$, and the explicit linear multistep integration method:*

$$\sum_{\sigma=0}^{k} \alpha_\sigma\, y_{n+\sigma} = \Delta t \sum_{\sigma=0}^{k-1} \beta_\sigma\, f_{n+\sigma}\,, \tag{2.7}$$

with the normalization $\alpha_k = 1$, $|\alpha_0| + |\beta_0| > 0$. Assume that the method is stable (in the sense of [HNW93, Definition III.3.2]*) and that it is of order 1 or higher. Then the stability region for this method contains no positive real number.*

Proof. Following [HNW93, HW96], we introduce the polynomials

$$\varrho(X) := \sum_{j=0}^{k} \alpha_j\, X^j\,, \quad \sigma(X) := \sum_{j=0}^{k-1} \beta_j\, X^j\,.$$

The assumptions of Lemma 2.4 can be rephrased as:

$$\varrho(1) = 0\,, \quad \varrho'(1) = \sigma(1) \neq 0\,,$$

and ϱ has no root of z satisfying $|z| > 1$. In particular, $\varrho'(1)$ must be positive for otherwise (recall $\alpha_k = 1$) ϱ would have a real root in the open interval $(1, +\infty)$. We therefore have $\sigma(1) > 0$.

For any given $\mu > 0$, the real polynomial:

$$P_\mu(X) := \varrho(X) - \mu\, \sigma(X)\,,$$

has degree k and is unitary. It tends to $+\infty$ at $+\infty$ and $P_\mu(1) = -\mu\,\sigma(1) < 0$. Hence P_μ vanishes in the open interval $(1, +\infty)$ and μ does not belong to the stability region of the numerical method. □

Lemma 2.4 is consistent with the plots in [HW96] of the stability regions for the explicit Adams and Nyström methods. Observe however that some stability regions may contain complex numbers of positive real part, e. g., the explicit Adams method of order 3.

Proof of Lemma 2.3. Under assumption 2.1, $\mathcal{A}$ has no other zero on $\mathbb{S}^1$ than $z = 1$ (with multiplicity 1). On the other hand, $\mathcal{A}$ admits a unique pole over $\mathbb{C}$, at $z = 0$ and of order r (because we have $a_{-r} \neq 0$). The cornerstone of the forthcoming proof is the residue theorem for meromorphic functions. Being given Γ a direct closed complex contour encircling the origin once and on which $\mathcal{A}$ does not vanish, then

$$\frac{1}{2\,i\,\pi} \int_\Gamma \frac{\mathcal{A}'(z)}{\mathcal{A}(z)}\, \mathrm{d}z = \#\{\text{zeros inside } \Gamma\} - \#\{\text{poles inside } \Gamma\}\,, \tag{2.8}$$

where zeros and poles are counted with multiplicity. The second integer on the right hand side equals r, and we intend now to compute $R_\Gamma := \#\{\text{zeros inside } \Gamma\}$ thanks to an appropriate choice for the contour Γ (for which $R_\Gamma = R$).

The contour Γ_ε. Let us consider some parameter $\varepsilon \in (0,\pi/4]$ sufficiently small (to be determined later on), and let us define the contour Γ_ε as $\mathbb{S}^1$ but for a small chord avoiding 1, see Figure 2.1. More precisely, we consider the path Γ_ε as the union $\Gamma_{\varepsilon,1} \cup \Gamma_{\varepsilon,2}$, with:

$$\Gamma_{\varepsilon,1} := \Big\{ \mathrm{e}^{i\,\theta} \, , \, \theta \in [\varepsilon, 2\,\pi-\varepsilon] \Big\} \, , \quad \Gamma_{\varepsilon,2} := \Big\{ \cos\varepsilon + i\,\omega \, , \, \omega \in [-\sin\varepsilon, \sin\varepsilon] \Big\} \, .$$

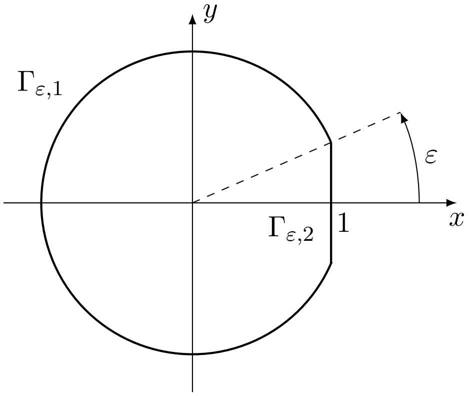

FIGURE 2.1. The integration contour Γ_ε.

Choice of the parameter ε. Let us first observe that 1 is a simple zero of $\mathcal{A}$ so we can choose $\varepsilon_0 > 0$ small enough such that, for any $\varepsilon \in (0,\varepsilon_0]$, the number of zeros of $\mathcal{A}$ inside Γ_ε equals the number of zeros of $\mathcal{A}$ in $\mathbb{D} \setminus \{0\}$.

Our goal now is to show that, for any sufficiently small $\varepsilon > 0$, there holds

$$\begin{cases} \mp \Im \mathcal{A}(\mathrm{e}^{\pm i\,\varepsilon}) > 0 \, , & \text{if} \quad a < 0 \, , \\ \pm \Im \mathcal{A}(\mathrm{e}^{\pm i\,\varepsilon}) > 0 \, , & \text{if} \quad a > 0 \, , \end{cases}$$

and for all $z \in \Gamma_{\varepsilon,2}$, $a\,\mathcal{A}(z) \notin \mathbb{R}^+$. These properties follow from studying the variation of the function $\Im \mathcal{A}(\cos\varepsilon + i\,\omega)$. Namely, we compute

$$\frac{\mathrm{d}}{\mathrm{d}\omega} \Im \mathcal{A}(\cos\varepsilon + i\,\omega) = \Re \mathcal{A}'(\cos\varepsilon + i\,\omega) = a + \Re \left(\mathcal{A}'(\cos\varepsilon + i\,\omega) - \mathcal{A}'(1) \right) .$$

Consequently, if we assume $a > 0$, then $(\omega \mapsto \Im \mathcal{A}(\cos\varepsilon + i\,\omega))$ is increasing on $[-\sin\varepsilon, \sin\varepsilon]$, while if we assume $a < 0$, then $(\omega \mapsto \Im \mathcal{A}(\cos\varepsilon + i\,\omega))$ is decreasing on $[-\sin\varepsilon, \sin\varepsilon]$. In any of these two cases, $\mathcal{A}(\cos\varepsilon + i\,\omega)$ is real for at most one value of ω.

We now observe that $\mathcal{A}(\cos\varepsilon)$ is real. In particular, for any $0 < |\omega| \le \sin\varepsilon$, $\mathcal{A}(\cos\varepsilon + i\,\omega)$ belongs to $\mathbb{C} \setminus \mathbb{R}$ and the sign property for $\Im \mathcal{A}(\mathrm{e}^{\pm i\,\varepsilon})$ is proved. Furthermore, using $\mathcal{A}'(1) = a$, we find that $\mathcal{A}(\cos\varepsilon)$ is positive if a is negative while $\mathcal{A}(\cos\varepsilon)$ is negative if a is positive. We thus have, provided that ε is sufficiently small, $a\,\mathcal{A}(z) \notin \mathbb{R}^+$ for any $z \in \Gamma_{\varepsilon,2}$. From now on, ε is fixed and the latter properties hold.

Application of the residue theorem. We denote hereafter $\log_-$ the principal complex logarithm with the usual branch cut along $\mathbb{R}_-$, and $\log_+$ the complex logarithm with a branch cut along $\mathbb{R}_+$.

For any $z \in \Gamma_{\varepsilon,1}$, one has $\mathcal{A}(z) \neq 0$ (thanks to assumption 2.1) and $\mathcal{A}(z) \notin \mathbb{R}^{-,*}$ (for otherwise the stability region of (2.7) would contain $-\lambda\,\mathcal{A}(z) \in \mathbb{R}^{+,*}$, which can not hold by Lemma 2.4). We can thus use $\log_-$ for computing the integral along $\Gamma_{\varepsilon,1}$, and we get

$$\frac{1}{2\,i\,\pi} \int_{\Gamma_{\varepsilon,1}} \frac{\mathcal{A}'(z)}{\mathcal{A}(z)} \, \mathrm{d}z = \frac{1}{2\,i\,\pi} \Big(\log_- \mathcal{A}(\mathrm{e}^{-i\,\varepsilon}) - \log_- \mathcal{A}(\mathrm{e}^{i\,\varepsilon}) \Big) \, .$$

The integral along $\Gamma_{\varepsilon,2}$ depends on the sign of a.

- Suppose $a<0$. Then we know that for all $z\in\Gamma_{\varepsilon,2}$, $\mathcal{A}(z)$ does not belong to $\mathbb{R}^-$. We can again use the $\log_-$ logarithm and derive
$$\frac{1}{2\,i\,\pi}\int_{\Gamma_{\varepsilon,2}}\frac{\mathcal{A}'(z)}{\mathcal{A}(z)}\,\mathrm{d}z=\frac{1}{2\,i\,\pi}\Big(\log_-\mathcal{A}(\mathrm{e}^{i\,\varepsilon})-\log_-\mathcal{A}(\mathrm{e}^{-i\,\varepsilon})\Big)\,.$$
Summing the two contributions over $\Gamma_{\varepsilon,1}$ and $\Gamma_{\varepsilon,2}$ we finally obtain
$$\frac{1}{2\,i\,\pi}\int_{\Gamma_\varepsilon}\frac{\mathcal{A}'(z)}{\mathcal{A}(z)}\,\mathrm{d}z=0\,,$$
and $R=r$.
- Suppose now $a>0$. Then we know that for all $z\in\Gamma_{\varepsilon,2}$, $\mathcal{A}(z)$ does not belong to $\mathbb{R}^+$. We use the $\log_+$ logarithm and derive
$$\frac{1}{2\,i\,\pi}\int_{\Gamma_{\varepsilon,2}}\frac{\mathcal{A}'(z)}{\mathcal{A}(z)}\,\mathrm{d}z=\frac{1}{2\,i\,\pi}\Big(\log_+\mathcal{A}(\mathrm{e}^{i\,\varepsilon})-\log_+\mathcal{A}(\mathrm{e}^{-i\,\varepsilon})\Big)\,.$$
Summing the two contributions over $\Gamma_{\varepsilon,1}$ and $\Gamma_{\varepsilon,2}$, we obtain
$$R-r=\frac{1}{2\,i\,\pi}\Big(\log_+\mathcal{A}(\mathrm{e}^{i\,\varepsilon})-\log_-\mathcal{A}(\mathrm{e}^{i\,\varepsilon})\Big)-\frac{1}{2\,i\,\pi}\Big(\log_+\mathcal{A}(\mathrm{e}^{-i\,\varepsilon})-\log_-\mathcal{A}(\mathrm{e}^{-i\,\varepsilon})\Big)\,.$$
The difference $\log_+-\log_-$ equals 0 on $\Omega_+:=\big\{z\in\mathbb{C}\,,\,\Im z>0\big\}$ and equals $2\,i\,\pi$ on $\Omega_-:=\big\{z\in\mathbb{C}\,,\,\Im z<0\big\}$. To complete the proof, we recall that $\mathcal{A}(\mathrm{e}^{\pm i\,\varepsilon})$ belong to $\Omega_\pm$, and we thus get $R-r=-1$.

□

2.3. **The leading boundary layer profile.** Using the flux function F, we can rewrite the boundary layer profile equations (2.3), (2.4), and introduce the following definition.

Definition 2.5. *Being given a real number u, we call $(v_j)_{j\in\mathbb{N}}$ a* boundary layer profile *associated with u a sequence that satisfies the following requirements:*

(i) $v_0=\cdots=v_{r-1}=-u$,
(ii) $F(u+v_j,\dots,u+v_{j+p+r-1})=F(u,\dots,u)$, *for all $j\geq 0$,*
(iii) $\lim_{j\to\infty}v_j=0$.

Let us comment some facts. The first point *(i)* above is related to the Dirichlet condition (2.4) with u in place of $u^{\mathrm{int}}(0,t^n)$ (here the time variable is frozen). As a consequence of the linearity of the numerical flux F, the above condition *(ii)* reads
$$\forall\,j\geq 0\,,\quad\sum_{\ell=-r}^{p-1}f_\ell\,v_{j+\ell+r}=0\,,$$
which is equivalent to
$$\forall\,j\geq 0\,,\quad\sum_{\ell=-r}^{p}a_\ell\,v_{j+\ell+r}=0\,,\tag{2.9}$$
if condition *(iii)* is satisfied. Boundary layer profiles are therefore the zero-limit solutions to the linear recurrence relation (2.9) for which the r first terms of the sequence coincide.

Definition 2.6. *The set of all the values u such that a stable boundary layer associated to u exists is denoted*
$$\mathcal{C}_{\mathrm{num}}=\Big\{u\in\mathbb{R},\,\exists\,v\in\mathbb{R}^{\mathbb{N}}\text{ boundary layer profile associated with }u\Big\}\,.$$

This definition is the same as in [DL88, GS97, CHG01]. The set $\mathcal{C}_{\mathrm{num}}$ encodes the so-called *residual boundary conditions* for (1.1) coming from the continuous limit $\Delta t\to 0$ in (1.4), (1.6). We are now ready to prove the following result that characterizes the boundary layer profiles for the scheme (1.4).

Proposition 2.7. *Under Assumptions 1.1, 1.2, 1.3 and 2.1, there holds:*

- *if $a > 0$, then $\mathcal{C}_{\mathrm{num}} = \{0\}$ and the unique boundary layer profile associated with 0 is the zero sequence ($v_j = 0$ for all $j \geq 0$);*
- *if $a < 0$, then $\mathcal{C}_{\mathrm{num}} = \mathbb{R}$ and for any $u \in \mathbb{R}$ there is a unique boundary layer profile $(v_j)_{j\in\mathbb{N}}$ associated with u, that decreases exponentially fast at infinity. We may write*

$$v_j = u\, w_j\,, \quad j \geq 0\,, \tag{2.10}$$

where $(w_j)_{j\in\mathbb{N}}$ denotes the boundary layer profile associated with $u = 1$.

Proof. As explained above, our goal is to determine the zero-limit solutions to the recurrence (2.9) that satisfy condition *(i)* in Definition 2.5. We thus look for the (stable) roots to the polynomial equation

$$\sum_{\ell=-r}^{p} a_\ell\, z^{\ell+r} = 0\,.$$

Since this polynomial does not vanish at zero, its roots in $\mathbb{D}$ coincide (with equal multiplicity) with the zeros of $\mathcal{A}$ in $\mathbb{D} \setminus \{0\}$. Lemma 2.3 precisely gives the number of such zeros.

The zero-limit solutions to the linear recurrence (2.9) are spanned by the sequences $Z^{(m)}$ ($m = 1, \dots, r$ if $a < 0$ and $m = 1, \dots, r-1$ if $a > 0$):

$$(j^\nu\, z_i^j)_{j\in\mathbb{N}}\,, \quad 0 \leq \nu < \mu_i\,,\ 1 \leq i \leq q\,, \tag{2.11}$$

where $z_1, \dots, z_q$ denote the pairwise distinct zeros of $\mathcal{A}$ in $\mathbb{D} \setminus \{0\}$ and $\mu_1, \dots, \mu_q$ their corresponding multiplicity.

- We first assume $a > 0$. The subspace of zero-limit solutions to the linear recurrence (2.9) has dimension $r - 1$. Let $u \in \mathbb{R}$. We are looking for a sequence $v = \sum_{m=1}^{r-1} \omega_m\, Z^{(m)}$ such that $v_0 = \cdots = v_{r-1} = -u$, which is equivalent to

$$\begin{pmatrix} Z_0^{(1)} & \dots & Z_0^{(r-1)} & 1 \\ \vdots & & \vdots & \vdots \\ Z_{r-1}^{(1)} & \dots & Z_{r-1}^{(r-1)} & 1 \end{pmatrix} \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_{r-1} \\ u \end{pmatrix} = 0\,.$$

  The involved matrix in $\mathcal{M}_{r,r}(\mathbb{C})$ is invertible and therefore $u = 0$, $v = 0$.
- We now assume $a < 0$. The subspace of zero-limit solutions to the linear recurrence (2.9) has dimension r. Let $u \in \mathbb{R}$. We are looking for a sequence $v = \sum_{m=1}^{r} \omega_m\, Z^{(m)}$ such that $v_0 = \cdots = v_{r-1} = -u$, which is equivalent to

$$\begin{pmatrix} Z_0^{(1)} & \dots & Z_0^{(r)} \\ \vdots & & \vdots \\ Z_{r-1}^{(1)} & \dots & Z_{r-1}^{(r)} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_r \end{pmatrix} = -u \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\,.$$

  The involved matrix of $\mathcal{M}_{r,r}(\mathbb{C})$ is invertible and thus, for each given $u \in \mathbb{R}$ there is a unique solution $(\omega_1, \dots, \omega_r) \in \mathbb{C}^r$ which determines the boundary layer profile associated with u. By linearity, this profile takes the form (2.10) and it is exponentially decreasing.

□

2.4. **The first boundary layer corrector.** Our goal in this subsection is to construct a solution to the first boundary layer corrector equations (2.5), (2.6). In what follows, the function $u^{\mathrm{bl},0}$ will be a boundary layer profile associated with some discretized trace of the exact solution to (1.1). In the case $a > 0$, there is no boundary layer profile but zero and the solution to (2.5), (2.6) is also zero. In the case $a < 0$, the space of boundary layer profiles is spanned by the sequence $(w_j)_{j\in\mathbb{N}}$, and it is therefore sufficient to construct a sequence that satisfies

$$\begin{aligned} &\forall\, j \geq r\,, \quad \sum_{\ell=-r}^{p} a_\ell\, \widetilde{w}_{j+\ell} + w_j = 0\,, \\ &\widetilde{w}_0 = \cdots = \widetilde{w}_{r-1} = 0\,, \quad \lim_{j \to \infty} \widetilde{w}_j = 0\,. \end{aligned} \tag{2.12}$$

Lemma 2.8. *Under the assumptions of Proposition 2.7, in the case $a < 0$, there exists a unique solution $(\widetilde{w}_j)_{j \in \mathbb{N}}$ to (2.12) and this solution decays exponentially fast at infinity.*

Proof. Uniqueness easily follows from the linearity of (2.12) and Proposition 2.7. As far as existence is concerned, we keep the notation of Proposition 2.7 and decompose the sequence $(w_j)_{j \in \mathbb{N}}$ as

$$w = \sum_{m=1}^{r} \omega_m\, Z^{(m)}\,,$$

where the $Z^{(m)}$'s are given by (2.11). Due to the linearity of (2.12), we first construct a zero-limit solution to the recurrence

$$\forall\, j \geq r\,, \quad \sum_{\ell=-r}^{p} a_\ell\, W^{(m)}_{j+\ell} + Z^{(m)}_j = 0\,, \quad Z^{(m)}_j = j^\nu\, z_i^j\,,$$

which is done by choosing $W^{(m)}$ of the form

$$W^{(m)}_j = \sum_{\mu=0}^{\mu_i - 1} \varsigma_\mu\, j^{\mu_i + \mu}\, z_i^j\,,$$

and by identifying the coefficients $\varsigma_0, \dots, \varsigma_{\mu_i - 1}$ (this procedure gives an invertible upper triangular system). Summing finitely many such sequences $W^{(m)}$, we get a sequence W that decays exponentially at infinity and that is a solution to the recurrence relation

$$\forall\, j \geq r\,, \quad \sum_{\ell=-r}^{p} a_\ell\, W_{j+\ell} + w_j = 0\,.$$

The sequence $(\widetilde{w}_j)$ is obtained by correcting the initial conditions for (W_j), that is by choosing

$$\widetilde{w} := W + \sum_{m=1}^{r} \varpi_m\, Z^{(m)}\,,$$

with

$$\begin{pmatrix} Z^{(1)}_0 & \dots & Z^{(r)}_0 \\ \vdots & & \vdots \\ Z^{(1)}_{r-1} & \dots & Z^{(r)}_{r-1} \end{pmatrix} \begin{pmatrix} \varpi_1 \\ \vdots \\ \varpi_r \end{pmatrix} = - \begin{pmatrix} W_0 \\ \vdots \\ W_{r-1} \end{pmatrix}.$$

□

2.5. **The approximate solution.** Let us recall that the solution to (1.1), supplemented with the homogeneous Dirichlet condition (1.2) in the case $a > 0$, is given by

$$u^{\text{ex}}(x,t) = u_0(x - a\,t), \quad x \geq 0, \quad t \geq 0\,, \tag{2.13}$$

where the initial condition u_0 has been extended by 0 to $\mathbb{R}^-$ in the case $a > 0$. This suggests defining the interior numerical solution as

$$u^{\text{int}}_{j,n} := \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} u_0(x - a\, t^n)\, \mathrm{d}x\,, \quad j \geq 0\,, \quad n \geq 0\,.$$

In particular, (1.7) gives

$$\forall\, n = 0, \dots, k-1\,, \quad \forall\, j \geq 0\,, \quad u^{\text{int}}_{j,n} = u^n_j\,.$$

In the case $a > 0$, there is no boundary layer and we define the approximate solution $u^{\rm app}$ to (1.4), (1.6) as

$$u^{\rm app}_{j,n} := u^{\rm int}_{j,n} \, , \quad j \geq 0 \, , \quad n \geq 0 \, .$$

In the case $a < 0$, there exists a one-dimensional space of boundary layer profiles and we can also construct boundary layer correctors. In view of (2.4), we first need to approximate the trace of the exact solution $u^{\rm ex}$ and therefore set

$$\forall \, n \geq 0 \, , \quad u^{\rm tr}_n := \frac{1}{\Delta t} \, \int_{t^n}^{t^{n+1}} u_0(-a \, t) \, {\rm d}t \, .$$

We now define the leading order boundary layer profile $u^{\rm bl,0}$ and first order boundary layer corrector $u^{\rm bl,1}$ as follows:

$$u^{\rm bl,0}_{j,n} := \begin{cases} 0 \, , & j \geq 0 \, , \quad n = 0,\dots,k-1 \, , \\ u^{\rm tr}_n \, w_j \, , & j \geq 0 \, , \quad n \geq k \, , \end{cases}$$

$$u^{\rm bl,1}_{j,n} := \begin{cases} 0 \, , & j \geq 0 \, , \quad n = 0,\dots,k-1 \, , \\ \left(\Delta t \, \sum_{\sigma=0}^{k-1} \beta_\sigma \right)^{-1} \, \left(\sum_{\sigma=0}^k \alpha_\sigma \, u^{\rm tr}_{n+\sigma} \right) \, \widetilde{w}_j \, , & j \geq 0 \, , \quad n \geq k \, . \end{cases}$$

The approximate solution $u^{\rm app}$ to (1.4), (1.6) is then defined by:

$$u^{\rm app}_{j,n} := u^{\rm int}_{j,n} + u^{\rm bl,0}_{j,n} + \Delta x \, u^{\rm bl,1}_{j,n} \, , \quad j \geq 0 \, , \quad n \geq 0 \, . \tag{2.14}$$

Thanks to our choice for the initial data, we again have:

$$u^{\rm app}_{j,n} = u_j^n \, , \quad j \geq 0 \, , \quad n = 0,\dots,k-1 \, .$$

3. Proof of the main result

The error analysis uses the expression of the approximate solution $(u^{\rm app}_{j,n})_{j \geq 0, n \geq 0}$ introduced in subsection 2.5. We thus focus on the interior consistency error that is defined by:

$$\varepsilon_{j,n+k} := \frac{1}{\Delta t} \, \left(\sum_{\sigma=0}^k \alpha_\sigma \, u^{\rm app}_{j,n+\sigma} + \lambda \, \sum_{\sigma=0}^{k-1} \beta_\sigma \, \sum_{\ell=-r}^p a_\ell \, u^{\rm app}_{j+\ell,n+\sigma} \right) \, ,$$

with $j \geq r$ and $n \geq 0$, and on the boundary errors:

$$\eta_{j,n} := u^{\rm app}_{j,n} \, , \quad 0 \leq j \leq r-1 \, , \quad n \geq k \, .$$

We recall that the approximate solution $u^{\rm app}$ has the same initial data as the exact numerical solution (whatever the sign of a):

$$u^{\rm app}_{j,n} = u_j^n \, , \quad j \geq 0 \, , \quad n = 0,\dots,k-1 \, .$$

Consequently, the error:

$$e_{j,n} := u^{\rm app}_{j,n} - u_j^n \, , \quad j \geq 0 \, , \quad n \geq 0 \, ,$$

is a solution to the following numerical scheme with presumably small forcing terms and zero initial data:

$$\begin{cases} \sum_{\sigma=0}^k \alpha_\sigma \, e_{j,n+\sigma} + \lambda \, \sum_{\sigma=0}^{k-1} \beta_\sigma \, \sum_{\ell=-r}^p a_\ell \, e_{j+\ell,n+\sigma} = \Delta t \, \varepsilon_{j,n+k} \, , & j \geq r \, , \quad n \geq 0 \, , \\ e_{j,n} = \eta_{j,n} \, , & 0 \leq j \leq r-1 \, , \quad n \geq k \, , \\ e_{j,0} = \cdots = e_{j,k-1} = 0 \, , & j \geq 0 \, . \end{cases} \tag{3.1}$$

The aim of the following two subsections is to quantify the smallness of the source terms in (3.1) in order to apply the stability estimate of [GT81]. The smallness of the source terms will yield, up to losing some powers of Δt, a semigroup estimate for $(e_{j,n})_{j \geq 0, n \geq 0}$ which will eventually give the semigroup estimate for the numerical solution $(u_j^n)_{j \geq 0, n \geq 0}$.

3.1. **The case of an incoming velocity.** We assume here $a > 0$ so that no boundary layer arises in the solution to (1.4), (1.6) ($\mathscr{C}_{\rm num} = \{0\}$). The approximate solution merely reads:

$$u^{\rm app}_{j,n} = \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} u_0(y - a\,t^n)\,{\rm d}y\,, \quad j \ge 0\,, \quad n \ge 0\,,$$

where we recall that u_0 has been extended by zero to $\mathbb{R}^-$. From the flatness conditions $u_0(0) = u_0'(0) = 0$, we have $u_0 \in H^2(\mathbb{R})$. The errors in (3.1) satisfy the following bounds.

Proposition 3.1. *Let us assume $a > 0$. Under the assumptions of Theorem 1.5 and in the CFL regime (1.5), there exists a constant $C > 0$ that is independent of u_0 and $\Delta t \in (0,1]$ such that*

$$\sup_{n\ge k} \sum_{j\ge r} \Delta x\,|\varepsilon_{j,n}|^2 \le C\,\Delta t^2\,\|u_0''\|^2_{L^2(\mathbb{R}^+)}\,, \tag{3.2}$$

$$\sum_{n\ge k} \sum_{j=0}^{r-1} \Delta t\,|\eta_{j,n}|^2 \le C\,\Delta t^2\,\|u_0'\|^2_{L^2(\mathbb{R}^+)}\,. \tag{3.3}$$

Proof. Let us first consider the boundary error terms $(\eta_{j,n})$. Since u_0 vanishes on $\mathbb{R}^-$, there holds $\eta_{j,n} = 0$ if $n \ge r/(a\,\lambda)$. The sum in (3.3) therefore reduces to finitely many terms (and the number of such terms is independent of Δt). We consider some space index $j \in \{0,\dots,r-1\}$ and some time index $k \le n < r/(a\,\lambda)$, and write

$$\eta_{j,n} = \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} u_0(x - a\,t^n)\,{\rm d}x = \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} \int_0^{x - a\,t^n} u_0'(y)\,{\rm d}y\,{\rm d}x\,.$$

We then apply the Cauchy-Schwarz inequality and get

$$|\eta_{j,n}|^2 \le C \int_{x_j}^{x_{j+1}} \left| \int_0^{x-a\,t^n} u_0'(y)^2\,{\rm d}y \right| {\rm d}x \le C\,\Delta t\,\|u_0'\|^2_{L^2(\mathbb{R}^+)}\,.$$

Summing the finitely many nonzero error terms, we get (3.3).

We now deal with the consistency error in the interior domain. Using the consistency assumptions 1.1 and 1.2, we have:

$$\begin{aligned}
\Delta t\,\varepsilon_{j,n+k} &= \frac{1}{\Delta x} \sum_{\sigma=0}^{k} \alpha_\sigma \int_{x_j}^{x_{j+1}} u_0(x - a\,t^{n+\sigma}) - u_0(x - a\,t^n)\,{\rm d}x \\
&+ \frac{\lambda}{\Delta x} \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} a_\ell \left(\int_{x_{j+\ell}}^{x_{j+\ell+1}} u_0(x - a\,t^{n+\sigma})\,{\rm d}x - \int_{x_j}^{x_{j+1}} u_0(x - a\,t^{n+\sigma})\,{\rm d}x \right) \\
&= -\frac{1}{\Delta x} \sum_{\sigma=0}^{k} \alpha_\sigma \int_{x_j}^{x_{j+1}} \int_{-a\,\sigma\,\Delta t}^{0} u_0'(x + y - a\,t^n)\,{\rm d}y\,{\rm d}x \\
&+ \frac{\lambda}{\Delta x} \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} a_\ell \int_{x_j}^{x_{j+1}} \int_0^{\ell\,\Delta x} u_0'(x + y - a\,t^{n+\sigma})\,{\rm d}y\,{\rm d}x \\
&= -\frac{\lambda}{\Delta x} \sum_{\sigma=0}^{k} \sigma\,\alpha_\sigma\,a \int_{x_j}^{x_{j+1}} \int_0^{\Delta x} u_0'(x - a\,\sigma\,\lambda\,y - a\,t^n)\,{\rm d}y\,{\rm d}x \\
&+ \frac{\lambda}{\Delta x} \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} \ell\,a_\ell \int_{x_j}^{x_{j+1}} \int_0^{\Delta x} u_0'(x + \ell\,y - a\,t^{n+\sigma})\,{\rm d}y\,{\rm d}x\,.
\end{aligned}$$

Using the consistency assumptions 1.1 and 1.2 again, we can add the zero quantity

$$\frac{\lambda}{\Delta x} \left(\sum_{\sigma=0}^{k} \sigma\,\alpha_\sigma\,a - \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} \ell\,a_\ell \right) \int_{x_j}^{x_{j+1}} \int_0^{\Delta x} u_0'(x - a\,t^n)\,{\rm d}y\,{\rm d}x\,,$$

and get

$$\begin{aligned}\Delta t\,\varepsilon_{j,n+k} = &\frac{\lambda}{\Delta x}\sum_{\sigma=0}^{k}\sigma\,\alpha_\sigma\,a\int_{x_j}^{x_{j+1}}\int_0^{\Delta x}\int_{-a\,\sigma\,\lambda\,y}^{0} u_0''(x+x'-a\,t^n)\,\mathrm{d}x'\,\mathrm{d}y\,\mathrm{d}x\\ &+\frac{\lambda}{\Delta x}\sum_{\sigma=0}^{k-1}\beta_\sigma\sum_{\ell=-r}^{p}\ell\,a_\ell\int_{x_j}^{x_{j+1}}\int_0^{\Delta x}\int_0^{\ell\,y-a\,\sigma\,\Delta t} u_0''(x+x'-a\,t^n)\,\mathrm{d}x'\,\mathrm{d}y\,\mathrm{d}x\,.\end{aligned}\tag{3.4}$$

We now apply successive Cauchy-Schwarz inequalities. In the CFL regime (1.5), we get for instance

$$\begin{aligned}\left|\int_{x_j}^{x_{j+1}}\int_0^{\Delta x}\int_{-a\,\sigma\,\lambda\,y}^{0} u_0''(x+x'-a\,t^n)\,\mathrm{d}x'\,\mathrm{d}y\,\mathrm{d}x\right|^2 &\le C\,\Delta t^3\int_{x_j}^{x_{j+1}}\int_0^{\Delta x}\int_{-a\,\sigma\,\lambda\,y}^{0} u_0''(x+x'-a\,t^n)^2\,\mathrm{d}x'\,\mathrm{d}y\,\mathrm{d}x\\ &\le C\,\Delta t^4\int_{x_j}^{x_{j+1}}\int_{-a\,\sigma\,\Delta t}^{0} u_0''(x+x'-a\,t^n)^2\,\mathrm{d}x'\,\mathrm{d}x\\ &\le C\,\Delta t^5\int_{x_j-a\,k\,\Delta t}^{x_{j+1}} u_0''(x-a\,t^n)^2\,\mathrm{d}x\,.\end{aligned}$$

The other error term in (3.4) is estimated similarly, and in the end, we can show that there exists a fixed integer $j_0>0$ (that only depends on the CFL number λ, a and k) such that

$$\forall\, j\ge r\,,\quad \forall\, n\in\mathbb{N}\,,\quad |\varepsilon_{j,n+k}|^2\le C\,\Delta t\int_{x_{j-j_0}}^{x_{j+p+1}} u_0''(x-a\,t^n)^2\,\mathrm{d}x\,.$$

The estimate (3.2) follows immediately. □

3.2. **The case of an outgoing velocity.** From now on, we consider the case of an outgoing velocity $a<0$ for which non-trivial boundary layers appear in the solution to the numerical scheme (1.4), (1.6) ($\mathfrak{C}_{\text{num}}=\mathbb{R}$). The following Proposition provides error bounds for the source terms in the numerical scheme (3.1).

Proposition 3.2. *Under the assumptions of Theorem 1.5 and in the CFL regime (1.5), there exists a constant $C>0$ that is independent of u_0 and $\Delta t\in(0,1]$ such that*

$$\sup_{n\ge 2\,k}\sum_{j\ge r}\Delta x\,|\varepsilon_{j,n}|^2\le C\,\Delta t^2\,\|u_0''\|_{L^2(\mathbb{R}^+)}^2\,,\tag{3.5}$$

$$\sup_{k\le n\le 2\,k-1}\sum_{j\ge r}\Delta x\,|\varepsilon_{j,n}|^2\le C\,\Delta t\,\|u_0'\|_{H^1(\mathbb{R}^+)}^2\,,\tag{3.6}$$

$$\sum_{n\ge k}\sum_{j=0}^{r-1}\Delta t\,|\eta_{j,n}|^2\le C\,\Delta t^2\,\|u_0'\|_{L^2(\mathbb{R}^+)}^2\,.\tag{3.7}$$

Proof. We first prove (3.7) and then deal with (3.5) and (3.6).
Errors at the boundary. We start with the proof of the estimate (3.7). From the definition (2.14), we obtain (recall $n\ge k$ and $j=0,\dots,r-1$):

$$\eta_{j,n}=u_{j,n}^{\text{app}}=\frac{1}{\Delta x}\int_{x_j}^{x_{j+1}}u_0(x+|a|\,t^n)\,\mathrm{d}x-\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}u_0(|a|\,t)\,\mathrm{d}t.$$

With the notation (1.5), the error $\eta_{j,n}$ can be written as

$$\begin{aligned}\eta_{j,n}&=\frac{\tau}{\Delta t}\int_0^{\Delta t/\tau}u_0(x_j+|a|\,t^n+|a|\,s)-u_0(|a|\,t^n)\,\mathrm{d}s-\frac{1}{\Delta t}\int_0^{\Delta t}u_0(|a|\,t^n+|a|\,s)-u_0(|a|\,t^n)\,\mathrm{d}s\\ &=\frac{\tau}{\Delta t}\int_0^{\Delta t/\tau}\int_0^{x_j+|a|\,s}u_0'(|a|\,t^n+y)\,\mathrm{d}y\,\mathrm{d}s-\frac{1}{\Delta t}\int_0^{\Delta t}\int_0^{|a|\,s}u_0'(|a|\,t^n+y)\,\mathrm{d}y\,\mathrm{d}s\,.\end{aligned}$$

Each term in $\eta_{j,n}$ is estimated by applying the Cauchy-Schwarz inequality. For instance, we have

$$\left| \frac{1}{\Delta t} \int_0^{\Delta t} \int_0^{|a|\,s} u_0'(|a|\,t^n + y)\,\mathrm{d}y\,\mathrm{d}s \right|^2 \le C\,\Delta t \int_0^{|a|\,\Delta t} u_0'(|a|\,t^n + y)^2\,\mathrm{d}y\,,$$

and similarly, we have

$$\left| \frac{\tau}{\Delta t} \int_0^{\Delta t/\tau} \int_0^{x_j + |a|\,s} u_0'(|a|\,t^n + y)\,\mathrm{d}y\,\mathrm{d}s \right|^2 \le C\,\Delta t \int_0^{r\,\Delta x} u_0'(|a|\,t^n + y)^2\,\mathrm{d}y\,.$$

Summing over the n's and the finitely many j's, we derive the bound (3.7).
Errors in the interior. We decompose the consistency error $\varepsilon_{j,n}$ in (3.1) as

$$\varepsilon_{j,n} = \varepsilon_{j,n}^{\mathrm{int}} + \varepsilon_{j,n}^{\mathrm{bl}}\,,$$

with self-explanatory notation. The estimate of the *interior* consistency error $\varepsilon_{j,n}^{\mathrm{int}}$ follows from the exact same arguments as we used in the case of an incoming velocity. The only difference is that, because of the sign of a, we do not need to extend u_0 by zero to $\mathbb{R}^-$ and no assumption on the behavior of u_0 at 0 is needed to derive the estimate

$$\sup_{n \ge k} \sum_{j \ge r} \Delta x\,|\varepsilon_{j,n}^{\mathrm{int}}|^2 \le C\,\Delta t^2\,\|u_0''\|_{L^2(\mathbb{R}^+)}^2\,. \tag{3.8}$$

We now focus on the new consistency error that comes from the boundary layer terms in u^{app}:

$$\begin{aligned}\varepsilon_{j,n+k}^{\mathrm{bl}} &= \frac{1}{\Delta t} \left(\sum_{\sigma=0}^{k} \alpha_\sigma\,u_{j,n+\sigma}^{\mathrm{bl},0} + \lambda \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} a_\ell\,u_{j+\ell,n+\sigma}^{\mathrm{bl},0} \right) + \frac{1}{\lambda} \sum_{\sigma=0}^{k} \alpha_\sigma\,u_{j,n+\sigma}^{\mathrm{bl},1} + \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} a_\ell\,u_{j+\ell,n+\sigma}^{\mathrm{bl},1} \\ &= \frac{1}{\Delta t} \sum_{\sigma=0}^{k} \alpha_\sigma\,u_{j,n+\sigma}^{\mathrm{bl},0} + \frac{1}{\lambda} \sum_{\sigma=0}^{k} \alpha_\sigma\,u_{j,n+\sigma}^{\mathrm{bl},1} + \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} a_\ell\,u_{j+\ell,n+\sigma}^{\mathrm{bl},1}\,.\end{aligned}$$

In the case $n \ge k$, we use the definition of the boundary layer profile and corrector $u^{\mathrm{bl},0}$, $u^{\mathrm{bl},1}$ to simplify the latter expression and get[5]

$$\varepsilon_{j,n+k}^{\mathrm{bl}} = \frac{1}{\lambda} \sum_{\sigma=0}^{k} \alpha_\sigma\,(u_{j,n+\sigma}^{\mathrm{bl},1} - u_{j,n}^{\mathrm{bl},1}) + \sum_{\sigma=0}^{k-1} \beta_\sigma \sum_{\ell=-r}^{p} a_\ell\,(u_{j+\ell,n+\sigma}^{\mathrm{bl},1} - u_{j+\ell,n}^{\mathrm{bl},1})\,.$$

The first boundary layer corrector is given in subsection 2.5. In particular, the error $\varepsilon_{j,n+k}^{\mathrm{bl}}$ can be decomposed as a linear combination of sequences of the form

$$\frac{\widetilde{w}_{j+\ell}}{\Delta t} \sum_{\sigma'=0}^{k} \alpha_{\sigma'}\,(u_{n+\sigma+\sigma'}^{\mathrm{tr}} - u_{n+\sigma'}^{\mathrm{tr}})\,,$$

with $\sigma = 0, \dots, k$ and $\ell = -r, \dots, p$. Since the sequence $(\widetilde{w}_j)$ is exponentially decreasing, we have

$$\begin{aligned}\sup_{n \ge k} \sum_{j \ge r} \Delta x\,|\varepsilon_{j,n+k}^{\mathrm{bl}}|^2 &\le C\,\Delta x\,\sup_{n \ge k} \sum_{\sigma=0}^{k} \frac{1}{\Delta t^2} \left| \sum_{\sigma'=0}^{k} \alpha_{\sigma'}\,(u_{n+\sigma+\sigma'}^{\mathrm{tr}} - u_{n+\sigma'}^{\mathrm{tr}}) \right|^2 \\ &\le \frac{C}{\Delta t} \sup_{n \ge k} \sum_{\sigma=0}^{k} \left| \sum_{\sigma'=0}^{k} \alpha_{\sigma'}\,(u_{n+\sigma+\sigma'}^{\mathrm{tr}} - u_{n+\sigma'}^{\mathrm{tr}} - u_{n+\sigma}^{\mathrm{tr}} + u_n^{\mathrm{tr}}) \right|^2\,.\end{aligned}$$

We compute

$$u_{n+\sigma+\sigma'}^{\mathrm{tr}} - u_{n+\sigma'}^{\mathrm{tr}} - u_{n+\sigma}^{\mathrm{tr}} + u_n^{\mathrm{tr}} = \frac{a^2}{\Delta t} \int_0^{\Delta t} \int_0^{\sigma\,\Delta t} \int_0^{\sigma'\,\Delta t} u_0''(|a|\,t^n + |a|\,s_1 + |a|\,s_2 + |a|\,s_3)\,\mathrm{d}s_3\,\mathrm{d}s_2\,\mathrm{d}s_1\,,$$

[5] Here we use the consistency assumption 1.2.

and the Cauchy-Schwarz inequality yields

$$|u^{\rm tr}_{n+\sigma+\sigma'}-u^{\rm tr}_{n+\sigma'}-u^{\rm tr}_{n+\sigma}+u^{\rm tr}_n|^2 \le C\,\Delta t^3 \int_0^{(2\,k+1)\,\Delta t} u_0''(|a|\,t^n+|a|\,s)^2\,{\rm d}s\,.$$

We have thus derived the estimate

$$\sup_{n\ge k} \sum_{j\ge r} \Delta x\,|\varepsilon^{\rm bl}_{j,n+k}|^2 \le C\,\Delta t^2\,\|u_0''\|^2_{L^2(\mathbb{R}^+)}\,.$$

Together with (3.8), this already proves (3.5).

We turn to the proof of (3.6). It only remains to estimate the ℓ^2_j norm of $(\varepsilon^{\rm bl}_{j,k}),\dots,(\varepsilon^{\rm bl}_{j,2\,k-1})$ because the interior errors $(\varepsilon^{\rm int}_{j,k}),\dots,(\varepsilon^{\rm int}_{j,2\,k-1})$ already satisfy (3.8), which is not larger than the right hand side in (3.6). Let us explain how we derive the estimate for $(\varepsilon^{\rm bl}_{j,k})$. The remaining terms are similar. The error $\varepsilon^{\rm bl}_{j,k}$ reads

$$\varepsilon^{\rm bl}_{j,k}=\frac{1}{\Delta t}\,u^{\rm bl,0}_{j,k}+\frac{1}{\lambda}\,u^{\rm bl,1}_{j,k}=\frac{w_j}{\Delta t}\,u^{\rm tr}_k+\frac{\widetilde{w_j}}{\lambda}\left(\Delta t\sum_{\sigma=0}^{k-1}\beta_\sigma\right)^{-1}\sum_{\sigma=0}^{k}\alpha_\sigma\,u^{\rm tr}_{k+\sigma}\,,$$

so we have

$$\sum_{j\ge r}\Delta x\,|\varepsilon^{\rm bl}_{j,k}|^2 \le \frac{C}{\Delta t}\sum_{\sigma=0}^{k}|u^{\rm tr}_{k+\sigma}|^2\,.$$

We now use the assumption $u_0(0)=0$ of Theorem 1.5 and get

$$u^{\rm tr}_{k+\sigma}=\frac{1}{\Delta t}\int_0^{\Delta t}\int_0^{|a|\,(t^{k+\sigma}+s)} u_0'(y)\,{\rm d}y\,{\rm d}s\,.$$

The Cauchy-Schwarz inequality then gives

$$|u^{\rm tr}_{k+\sigma}|^2 \le C\,\Delta t\int_0^{|a|\,t^{k+\sigma+1}} u_0'(y)^2\,{\rm d}y \le C\,\Delta t^2\,\|u_0'\|^2_{L^\infty(\mathbb{R}^+)} \le C\,\Delta t^2\,\|u_0'\|^2_{H^1(\mathbb{R}^+)}\,.$$

We thus get (3.6) for the sequence $(\varepsilon_{j,k})$ and the remaining terms $(\varepsilon^{\rm bl}_{j,k+1}),\dots,(\varepsilon^{\rm bl}_{j,2\,k-1})$ are dealt with in the same (rather crude) way. □

Propositions 3.1 and 3.2 imply the following result which uses GKS type norms.

Proposition 3.3. *Under the assumptions of Theorem 1.5 and in the CFL regime (1.5), there exists a constant $C>0$ that is independent of u_0 and $\Delta t\in(0,1]$, such that for all $\gamma>0$ there holds*

$$\sum_{n\ge k}\sum_{j\ge r}\Delta t\,\Delta x\,{\rm e}^{-2\,n\,\gamma\,\Delta t}\,|\varepsilon_{j,n}|^2 \le C\left(1+\frac{1}{\gamma}\right)\Delta t^2\,\|u_0\|^2_{H^2(\mathbb{R}^+)}\,, \tag{3.9}$$

$$\sum_{n\ge k}\sum_{j=0}^{r-1}\Delta t\,{\rm e}^{-2\,n\,\gamma\,\Delta t}\,|\eta_{j,n}|^2 \le C\,\Delta t^2\,\|u_0\|^2_{H^1(\mathbb{R}^+)}\,. \tag{3.10}$$

Proof of Proposition 3.3. The proof of (3.10) is immediate and follows from either (3.3) or (3.7) by using $\gamma>0$ (so that the exponential factors in (3.10) are not larger than 1).

The proof of (3.9) follows from either (3.2) or (3.5)-(3.6). In the incoming case ($a>0$), we use (3.2) and get

$$\begin{aligned}\sum_{n\ge k}\sum_{j\ge r}\Delta t\,\Delta x\,{\rm e}^{-2\,n\,\gamma\,\Delta t}\,|\varepsilon_{j,n}|^2 &\le C\,\Delta t^3\,\|u_0\|^2_{H^2(\mathbb{R}^+)}\sum_{n\ge k}{\rm e}^{-2\,n\,\gamma\,\Delta t}\\ &\le \frac{C}{{\rm e}^{2\,\gamma\,\Delta t}-1}\,\Delta t^3\,\|u_0\|^2_{H^2(\mathbb{R}^+)} \le \frac{C}{\gamma}\,\Delta t^2\,\|u_0\|^2_{H^2(\mathbb{R}^+)}\,,\end{aligned}$$

which is even better than (3.9). In the outgoing case, we use (3.5)-(3.6) and get

$$\sum_{n\geq k}\sum_{j\geq r}\Delta t\,\Delta x\,\mathrm{e}^{-2\,n\,\gamma\,\Delta t}\,|\varepsilon_{j,n}|^2 \leq C\,\Delta t^2\,\|u_0\|_{H^2(\mathbb{R}^+)}^2 + C\,\Delta t^3\,\|u_0\|_{H^2(\mathbb{R}^+)}^2 \sum_{n\geq 2\,k}\mathrm{e}^{-2\,n\,\gamma\,\Delta t}$$
$$\leq C\left(1+\frac{1}{\gamma}\right)\Delta t^2\,\|u_0\|_{H^2(\mathbb{R}^+)}^2\,.$$

□

Remark 3.4. *If we had not included the boundary layer corrector $u^{\mathrm{bl},1}$ in the approximate solution, the right hand side in the error estimate* (3.9) *would have been of the form $\Delta t\,\|u_0\|_{H^2(\mathbb{R}^+)}^2$ instead of $\Delta t^2\,\|u_0\|_{H^2(\mathbb{R}^+)}^2$, which would have not been sufficient to derive* (1.16) *because there is a loss of a factor Δt in the derivation of the estimate* (3.12) *below.*

3.3. **The semigroup estimate.** We now prove Theorem 1.5. We apply the main result[6] of [GT81] which states that the numerical scheme (3.1) is strongly stable in the sense of [GKS72]. In other words, there exists a constant $C > 0$, that is independent of the parameter $\gamma > 0$, such that there holds:

$$\frac{\gamma}{1+\gamma\,\Delta t}\sum_{n\geq 0}\sum_{j\geq 0}\Delta t\,\Delta x\,\mathrm{e}^{-2\,n\,\gamma\,\Delta t}\,|e_j^n|^2 + \sum_{n\geq 0}\sum_{j=0}^{r+p-1}\Delta t\,\mathrm{e}^{-2\,n\,\gamma\,\Delta t}\,|e_j^n|^2$$
$$\leq C\left(\frac{1+\gamma\,\Delta t}{\gamma}\sum_{n\geq k}\sum_{j\geq r}\Delta t\,\Delta x\,\mathrm{e}^{-2\,n\,\gamma\,\Delta t}\,|\varepsilon_j^n|^2 + \sum_{n\geq k}\sum_{j=0}^{r-1}\Delta t\,\mathrm{e}^{-2\,n\,\gamma\,\Delta t}\,|\eta_j^n|^2\right)$$
$$\leq C\,\Delta t^2\,\|u_0\|_{H^2(\mathbb{R}^+)}^2\left(\frac{\gamma\Delta t+1}{\gamma}\left(1+\frac{1}{\gamma}\right)+1\right)\,, \tag{3.11}$$

where we have used Proposition 3.3 to derive the second inequality in (3.11). We choose $\gamma = \Delta t^\mu$, with $\mu \in [0, 1/3]$. We thus derive from (3.11) the bound

$$\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,n\,\Delta t^{1+\mu}}\sum_{j\geq 0}\Delta x\,|e_j^n|^2 \leq C\,\Delta t^{2-3\,\mu}\,\|u_0\|_{H^2(\mathbb{R}^+)}^2\,.$$

In particular, a very crude lower bound for the left hand side gives

$$\sup_{n\geq 0}\,\mathrm{e}^{-2\,n\,\Delta t^{1+\mu}}\sum_{j\geq 0}\Delta x\,|e_j^n|^2 \leq C\,\Delta t^{1-3\,\mu}\,\|u_0\|_{H^2(\mathbb{R}^+)}^2\,. \tag{3.12}$$

The semigroup estimate (3.12) yields the bound

$$\forall\, n\in\mathbb{N}\,,\quad \sum_{j\geq 0}\Delta x\,|u_j^n|^2 \leq 2\sum_{j\geq 0}\Delta x\,|u_{j,n}^{\mathrm{app}}|^2 + C\,\mathrm{e}^{2\,n\,\Delta t^{1+\mu}}\,\Delta t^{1-3\,\mu}\,\|u_0\|_{H^2(\mathbb{R}^+)}^2\,,$$

with a constant C that is uniform with respect to all the parameters. We now derive a semigroup estimate for the approximate solution u^{app}. In the case of an incoming transport equation ($a > 0$), we have

$$u_{j,n}^{\mathrm{app}} = u_{j,n}^{\mathrm{int}} = \frac{1}{\Delta x}\int_{x_j}^{x_{j+1}} u_0(x-a\,t^n)\,\mathrm{d}x\,,$$

for all $j, n \in \mathbb{N}$ (recall that u_0 vanishes on $\mathbb{R}^-$). In particular, the Cauchy-Schwarz inequality yields

$$\sum_{j\geq 0}\Delta x\,|u_{j,n}^{\mathrm{app}}|^2 \leq \|u_0\|_{L^2(\mathbb{R}^+)}^2\,,$$

and we get

$$\forall\, n\in\mathbb{N}\,,\quad \sum_{j\geq 0}\Delta x\,|u_j^n|^2 \leq 2\,\|u_0\|_{L^2(\mathbb{R}^+)}^2 + C\,\Delta t^{1-3\,\mu}\,\mathrm{e}^{2\,n\,\Delta t^{1+\mu}}\,\|u_0\|_{H^2(\mathbb{R}^+)}^2\,,$$

[6] As a matter of fact, the main result of [GT81] requires more restrictive conditions than Assumption 1.3, but the extension of the result of [GT81] to numerical schemes that satisfy Assumption 1.3 was performed in [Cou13].

which gives (1.16). We now consider the case of an outgoing transport equation ($a<0$) and derive a semigroup estimate for the approximate solution $u^{\rm app}$. We still have

$$\sum_{j\geq 0}\Delta x\,|u^{\rm int}_{j,n}|^2\leq \|u_0\|^2_{L^2(\mathbb{R}^+)}\,,$$

and we thus focus on the semigroup estimate for the boundary layer profile and corrector. Let us first consider the boundary layer profile $u^{{\rm bl},0}$, for which we have

$$\begin{aligned}\sup_{n\geq 0}\sum_{j\geq 0}\Delta x\,|u^{{\rm bl},0}_{j,n}|^2=\sup_{n\geq k}\sum_{j\geq 0}\Delta x\,|u^{{\rm bl},0}_{j,n}|^2&=\sup_{n\geq k}\,\Delta x\,|u^{\rm tr}_n|^2\sum_{j\geq 0}w_j^2\\&\leq\sup_{n\geq k}\frac{C}{\Delta t}\left|\int_{t^n}^{t^{n+1}}u_0(|a|\,t)\,{\rm d}t\right|^2\leq C\,\|u_0\|^2_{L^2(\mathbb{R}^+)}\,.\end{aligned}$$

We now deal with the first boundary layer corrector $\Delta x\,u^{{\rm bl},1}$, for which we have

$$\begin{aligned}\sup_{n\geq 0}\sum_{j\geq 0}\Delta x\,|\Delta x\,u^{{\rm bl},1}_{j,n}|^2=\sup_{n\geq k}\sum_{j\geq 0}\Delta x^3\,|u^{{\rm bl},1}_{j,n}|^2&=\sup_{n\geq k}\,C\,\Delta x\left|\sum_{\sigma=0}^{k}\alpha_\sigma\,u^{\rm tr}_{n+\sigma}\right|^2\sum_{j\geq 0}\widetilde{w}_j^2\\&\leq C\,\Delta t\,\sup_{n\geq k}\sum_{\sigma=0}^{k}|u^{\rm tr}_{n+\sigma}|^2\leq C\,\|u_0\|^2_{L^2(\mathbb{R}^+)}\,.\end{aligned}$$

As in the incoming case, we have thus derived the bound

$$\sum_{j\geq 0}\Delta x\,|u^{\rm app}_{j,n}|^2\leq C\,\|u_0\|^2_{L^2(\mathbb{R}^+)}\,,$$

and we get (1.16) accordingly.

4. Example and counterexample

4.1. A 4 time-step 5 point centered scheme. As a first numerical illustration of the above results in the case of an outgoing velocity $a=-1$, we consider the following numerical scheme. The time-stepping is solved using the 3rd order explicit Adams-Bashforth method, so that assumption 1.2 is satisfied. The space discretization of the advection term $a\,\partial_x u$ is based on a centered five-point approximation supplemented with a fourth order stabilizing dissipative term:

$$\begin{aligned}u_j^{n+1}&=u_j^n-\lambda\left(\frac{23}{12}f_j^n-\frac{16}{12}f_j^{n-1}+\frac{5}{12}f_j^{n-2}\right)\,,\\f_j^n&:=a\,\frac{-u^n_{j+2}+8u^n_{j+1}-8u^n_{j-1}+u^n_{j-2}}{12}-\frac{-u^n_{j+2}+4u^n_{j+1}-6u^n_j+4u^n_{j-1}-u^n_{j-2}}{24}\,.\end{aligned}\tag{4.1}$$

As we will show with numerical experiments, this scheme displays numerical boundary layers when combined with Dirichlet boundary conditions. Let us compute $\mathcal{A}$:

$$\mathcal{A}(z)=\frac{a}{12}(-z^2+8z-8z^{-1}+z^{-2})-\frac{1}{24}(-z^2+4z-6+4z^{-1}-z^{-2}),$$

from which we get $\mathcal{A}(1)=0$ and $\mathcal{A}'(1)=a$ and the space discretization satisfies assumption 1.1. Moreover, for any $\theta\in\mathbb{R}$, one gets

$$\mathcal{A}({\rm e}^{i\theta})=-a\frac{i}{6}(\sin(2\theta)-8\sin(\theta))-\frac{1}{12}(-\cos(2\theta)+4\cos(\theta)-3)\,,\text{ and }\Re(\mathcal{A}(e^{i\theta}))=\frac{2}{3}\sin^4\left(\frac{\theta}{2}\right)\,,$$

Therefore the only root of $\mathcal{A}(e^{i\theta})$ in $[-\pi,\pi]$ is $\theta=0$. This ensures that assumption 2.1 is satisfied. Figure 4.1 below pictures the closed curve $\left\{-\lambda\mathcal{A}(e^{i\eta}),\eta\in\mathbb{R}\right\}$ for the choice $\lambda=0.4$ (blue curve), together with the stability domain of the time integrator (red dashed curve) ; see [HW96, HNW93] for details. Let us observe that the stability assumption for the Cauchy problem 1.3 is satisfied.

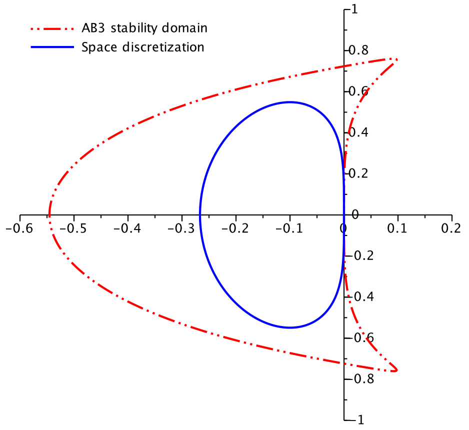

FIGURE 4.1. Verification of the stability assumption 1.3 for the 3rd order scheme (4.1).

The numerical test case concerns the following initial condition:

$$u_0(x) = \mathrm{e}^{-100(x-0.5)^2}, \quad x \in [0,1] ,$$

together with homogeneous Dirichlet conditions at both left and right boundaries (with no significant effect arising from the right boundary due to the incoming situation with $a = -1$ and the absence of boundary layer at $x = 1$). We compute the solution on $N = 216$ uniformly spaced grid cells, until time $T = 0.5$. At this time, the initial bump crosses the left boundary with the highest strength. As expected, the numerical solution (u_j^n) develops some boundary layer in the neighborhood of $x = 0$ due to the incompatibility of the homogeneous Dirichlet condition $u_j^n = 0$, $0 \leq j \leq r-1$, with the effective trace of the solution $u^{\rm int}(0^+, T) = 1$. We then observe on Figure 4.2 an oscillating pattern that does not disappear as Δx tends to 0. The two roots of $\mathcal{A}$ in $\mathbb{D} \setminus \{0\}$ are real and distinct; one of them equals approximately 0.0809 and therefore belongs to $(0,1)$, while the second one equals approximately -0.6595 and therefore belongs to $(-1,0)$, which gives rise to the oscillations in the boundary layer.

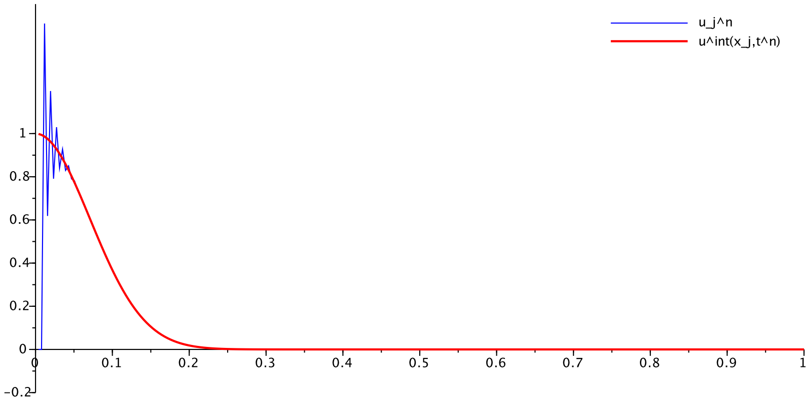

FIGURE 4.2. Numerical solution and exact solution at time T=0.5 (216 grid points).

The main term of the boundary layer expansion is a linear combination of two geometric sequences generated by the roots of the equation $\mathcal{A}(z) = 0$ in $\mathbb{D} \setminus \{0\}$ (see Lemma 2.3). In the present case, we obtain numerically $z_1 \simeq -0.6595$ and $z_2 \simeq 0.0809$. The precise boundary layer expansion $u^{\rm bl,0}(j,T) + \Delta x\, u^{\rm bl,1}(j,T)$ is depicted with crosses on the left picture of Figure 4.3 for the first 20 grid cells. Notice that it depends only on the trace of the solution at the considered time $u_n^{\rm tr}$ and on the discrete in time derivative of this trace, through $u^{\rm bl,1}$. It fits quite well the difference between the numerical solution and the exact one $u_j^n - u^{\rm int}(x_j,T)$. On the right picture of Figure 4.3 is represented the error in this boundary layer expansion $u_j^n - \big[u^{\rm int}(x_j,T) + u^{\rm bl,0}(j,T) + \Delta x\, u^{\rm bl,1}(j,T)\big]$ in the first 50 grid cells.

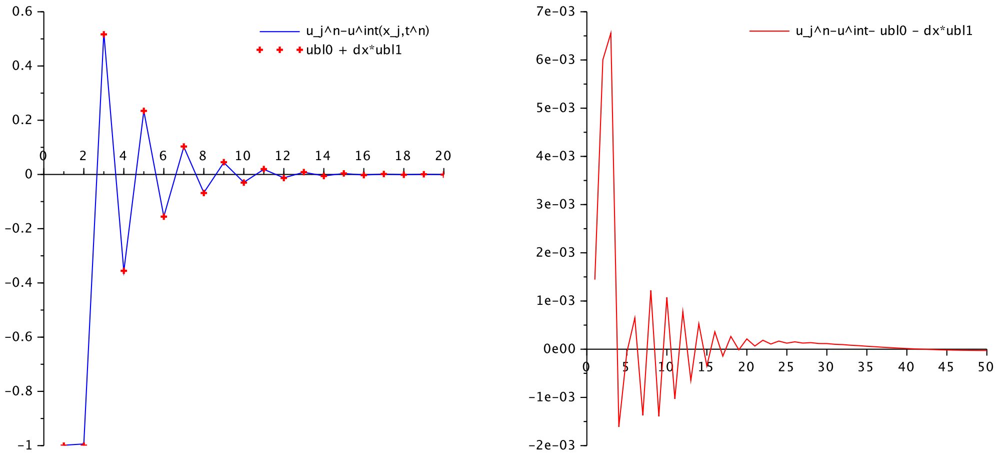

Figure 4.3. Boundary layer expansion at time T=0.5 (216 grid points)

The scheme (4.1) is third order in time and space accurate. We now consider the effective accuracy of this scheme for the IBVP problem by computing the $\ell^2([0,1])$ error at a final given time for successive values of 2^M grid points, $5 \le M \le 12$. More precisely, given a time $T > 0$, we compute the following two quantities, where $n = N_T$ is the first integer such that $N_T \Delta t \ge T$:

$$\left(\sum_{j=0}^{2^M} \Delta x \, \left| u_j^n - u^{\rm int}(x_j,t^n) \right|^2 \right)^{1/2} , \quad \text{and} \quad \left(\sum_{j=0}^{2^M} \Delta x \, \left| u_j^n - u^{\rm app}(x_j,t^n) \right|^2 \right)^{1/2} .$$

At a first time $T = 0.125$ at which no significant boundary layer has appeared at $x = 0$, the convergence of both quantities occur with order 3, see Figure 4.4 on the left. For very thin grids, one observes however a slight loss of accuracy when computing the usual numerical error. It corresponds to the presence of a very small boundary layer that deteriorates the effective order of accuracy.

At a later time $T = 0.4$ at which the boundary layer is sufficiently high to affect the convergence, the usual numerical error is strongly increased and the apparent order of accuracy is severely damaged: in Figure 4.4 on the right, we observe a numerical accuracy of order 0.5 for the usual numerical error, and of 1.5 for the error in the boundary layer expansion.

4.2. **The leap-frog scheme.** We now consider the usual three-time step leap-frog scheme, with a three point stencil in space:

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + a\, \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0\,. \tag{4.2}$$

The scheme corresponds to the so-called Nyström method of order 2 (also called the mid-point formula) combined with the center differentiation formula for the space discretization. The corresponding function $\mathcal{A}$ equals $z - z^{-1}$, and therefore vanishes at -1. Assumption 2.1 is no longer satisfied and Figure 4.5 below illustrates that the failure of Assumption 2.1 gives rise to a completely different behavior. Namely, we compute the numerical solution for (4.2) with $a = -1$ and homogeneous Dirichlet boundary conditions at different time levels, for the same kind of bump initial data. As the bump crosses the left boundary, a highly oscillatory wave packet emerges from the boundary and propagates with velocity $+1$ towards the right. The envelope of this wave packet is exactly the one of the initial condition, see Figure 4.5. The latter phenomenon has long been identified of course, see, e. g., [Tre82].

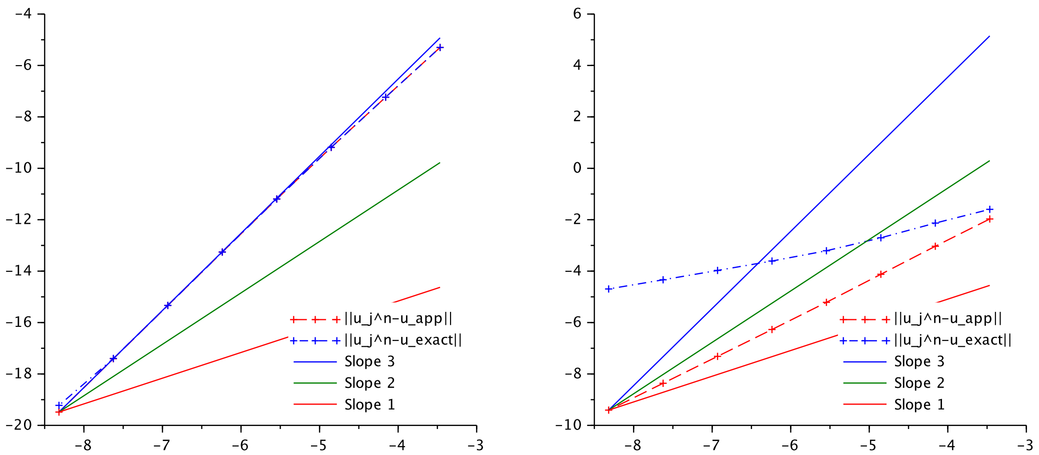

FIGURE 4.4. Convergence in log/log scale. Solution at time $T = 0.125$ with no significant boundary layer (left) / at time $T = 0.4$ with an important boundary layer (right).

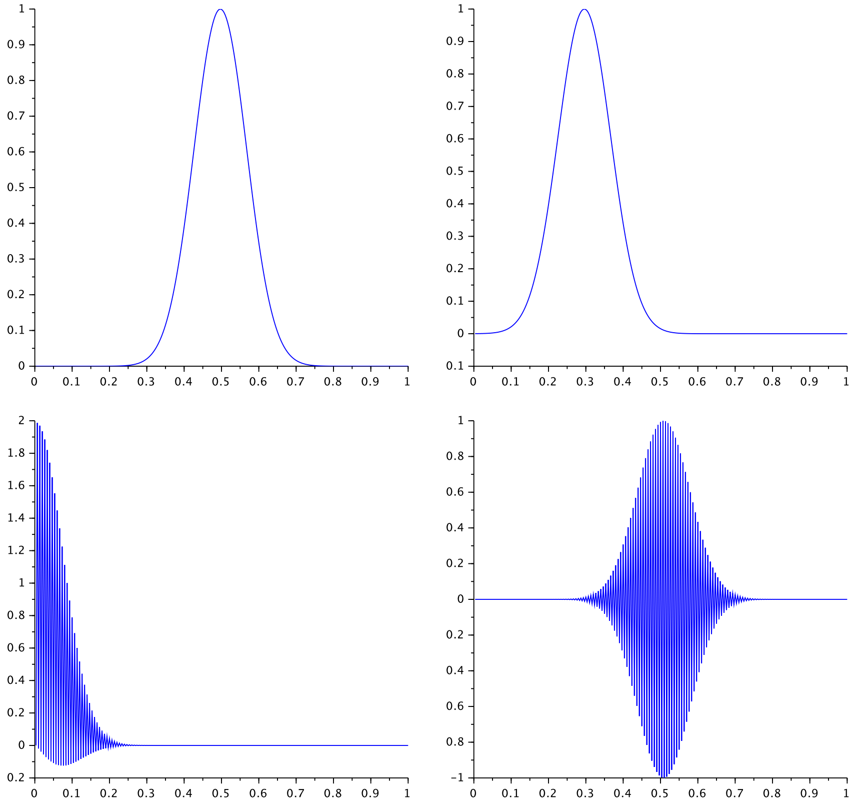

FIGURE 4.5. Leap-frog scheme, solution at time $T = 0$, $T = 0.2$, $T = 0.5$ and $T = 1$

REFERENCES

[CFL28] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100(1):32–74, 1928.

[CG11] J.-F. Coulombel and A. Gloria. Semigroup stability of finite difference schemes for multidimensional hyperbolic initial boundary value problems. *Math. Comp.*, 80(273):165–203, 2011.

[CHG01] C. Chainais-Hillairet and E. Grenier. Numerical boundary layers for hyperbolic systems in 1-D. *M2AN Math. Model. Numer. Anal.*, 35(1):91–106, 2001.

[Cou13] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems. In *HCDTE Lecture Notes. Part I. Nonlinear Hyperbolic PDEs, Dispersive and Transport Equations*, pages 97–225. American Institute of Mathematical Sciences, 2013.

[DL88] F. Dubois and P. LeFloch. Boundary conditions for nonlinear hyperbolic systems of conservation laws. *J. Differential Equations*, 71(1):93–122, 1988.

[GKO95] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods.* John Wiley & Sons, 1995.

[GKS72] B. Gustafsson, H.-O. Kreiss, and A. Sundström. Stability theory of difference approximations for mixed initial boundary value problems. II. *Math. Comp.*, 26(119):649–686, 1972.

[GS97] M. Gisclon and D. Serre. Conditions aux limites pour un système strictement hyperbolique fournies par le schéma de Godunov. *RAIRO Modél. Math. Anal. Numér.*, 31(3):359–380, 1997.

[GT81] M. Goldberg and E. Tadmor. Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comp.*, 36(154):603–626, 1981.

[HNW93] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I.* Springer-Verlag, second edition, 1993. Nonstiff problems.

[HW96] E. Hairer and G. Wanner. *Solving ordinary differential equations. II.* Springer-Verlag, second edition, 1996. Stiff and differential-algebraic problems.

[Kre70] H.-O. Kreiss. Initial boundary value problems for hyperbolic systems. *Comm. Pure Appl. Math.*, 23:277–298, 1970.

[Mét14] G. Métivier. On the L^2 well-posedness of hyperbolic initial boundary value problems. *Preprint*, 2014.

[TE05] L. N. Trefethen and M. Embree. *Spectra and pseudospectra.* Princeton University Press, 2005. The behavior of nonnormal matrices and operators.

[Tre82] L. N. Trefethen. Group velocity in finite difference schemes. *SIAM Rev.*, 24(2):113–136, 1982.

[Wu95] L. Wu. The semigroup stability of the difference approximations for initial-boundary value problems. *Math. Comp.*, 64(209):71–88, 1995.

IRMAR (UMR CNRS 6625), Université de Rennes, Campus de Beaulieu, 35042 Rennes Cedex, France.
E-mail address: Benjamin.Boutin@univ-rennes1.fr

CNRS, Université de Nantes, Laboratoire de Mathématiques Jean Leray (CNRS UMR6629), 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France.
E-mail address: Jean-Francois.Coulombel@univ-nantes.fr